Colombia's cyberinfrastructure for biodiversity: Building data infrastructure in emerging countries to foster socioeconomic growth

de Vega, Jose; Davey, Robert P.; Duitama, Jorge; Escobar, Dairo; Cristancho-ardila, Marco A.; Etherington, Graham J.; Minotto, Alice; Arenas-Suarez, Nelson E.; Pineda-Cardenas, Juan D.; Correa-Alvarez, Javier; Camargo Rodriguez, Anyela V.; Haerty, Wilfried; Mallarino-Robayo, Juan P.; Barreto-Hernandez, Emiliano; Muñoz-Torres, Monica; Fernandez Fuentes, Narcis; Di Palma, Federica

Published in:
Plants, People, Planet

DOI:
10.1002/ppp3.10086

Publication date:
2019

Citation for published version (APA):
de Vega, J., Davey, R. P., Duitama, J., Escobar, D., Cristancho-ardila, M. A., Etherington, G. J., Minotto, A., Arenas-Suarez, N. E., Pineda-Cardenas, J. D., Correa-Alvarez, J., Camargo Rodriguez, A. V., Haerty, W., Mallarino-Robayo, J. P., Barreto-Hernandez, E., Muñoz-Torres, M., Fernandez Fuentes, N., & Di Palma, F. (2019). Colombia's cyberinfrastructure for biodiversity: Building data infrastructure in emerging countries to foster socioeconomic growth. *Plants, People, Planet*. https://doi.org/10.1002/ppp3.10086

Received: 20 September 2019 | Revised: 20 September 2019 | Accepted: 2 October 2019

DOI: 10.1002/ppp3.10086

BRIEF REPORT

Plants People Planet PPP Open Access

Colombia's cyberinfrastructure for biodiversity: Building data infrastructure in emerging countries to foster socioeconomic growth

Jose J. De Vega[1] | Robert P. Davey[1] | Jorge Duitama[2] | Dairo Escobar[3] | Marco A. Cristancho-Ardila[4] | Graham J. Etherington[1] | Alice Minotto[1] | Nelson E. Arenas-Suarez[5] | Juan D. Pineda-Cardenas[6] | Javier Correa-Alvarez[7] | Anyela V. Camargo Rodriguez[8] | Wilfried Haerty[1] | Juan P. Mallarino-Robayo[4] | Emiliano Barreto-Hernandez[9] | Monica Muñoz-Torres[10] | Narcis Fernandez-Fuentes[11] | Federica Di Palma[1] | the Colombian Cyberinfrastructure Consortium for Biodiversity*

[1]Earlham Institute, Norwich Research Park, Norwich, UK

[2]Systems and Computing Engineering Department, Universidad de Los Andes, Bogotá, Colombia

[3]SiB Colombia - Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Bogotá, Colombia

[4]Faculty of Sciences, Universidad de Los Andes, Bogotá, Colombia

[5]Faculty of Sciences, Universidad Antonio Nariño, Bogotá, Colombia

[6]Apolo Supercomputing Centre, EAFIT University, Medellín, Colombia

[7]Faculty of Sciences, EAFIT University, Medellín, Colombia

[8]The John Bingham Laboratory, NIAB, Cambridge, UK

[9]Biotechnology Institute, Universidad Nacional de Colombia, Bogotá, Colombia

[10]Translational and Integrative Sciences Lab, Oregon State University, Corvallis, OR, USA

[11]IBERS, Aberystwyth University, Aberystwyth, UK

Correspondence
Jose J. De Vega, Earlham Institute, Norwich Research Park, NR4 7UZ, UK.
Email: jose.devega@earlham.ac.uk

Funding information
UK Research and Innovation (UKRI) Global Challenges Research Fund (GCRF), Grant/Award Number: BB/P028098/1

Societal Impact Statement

Colombia is a "megadiverse" country with vast natural resources. A history of recent conflict means that information is only now being collected on the natural capital of regions that were previously unexplored. Better access to data, tools, and expertise is required for evidence-supported decisions on the conservation of these resources. The development of a bespoke cyberinfrastructure could help fulfill this need by providing access to digital resources in a collaborative cyberenvironment. We outline key priorities and develop a reference framework for building cyberinfrastructure in Colombia. This framework could be applied to other fields and countries to promote knowledge exchange, scientific innovation, and socioeconomic growth.

KEYWORDS

biodiversity, data management, e-infrastructure, innovation, research policy

*Additional members of C3Biodiversidad, the Colombian Cyberinfrastructure Consortium for Biodiversity, are listed in the acknowledgements.

1 | DATA-DRIVEN INNOVATION AND DEVELOPMENT

Science and innovation are not a luxury but a prerequisite for social and economic development (Annan, 2003). Across different fields, acquisition and analysis of large amounts of data have become a common practice to drive innovation (Yang, Huang, Li, Liu, & Hu, 2017), particularly with today's highly instrumented data collection methods (Borgman, Wallis, & Mayernik, 2012). The efficient analysis of such data has an unprecedented potential to transform how we tackle the major challenges faced by humanity, from climate change to food security (Hilbert, 2016).

Data-driven innovation can only be achieved through greater access to data, through effective and efficient-enabling resources, and ensuring that the best available expertise is harnessed through them. This is particularly the case when collaboration is needed to address the research questions at a continental scale, such as the effect of global impacts on rich, vast ecological systems in the present climate change scenario (Peters, Loescher, SanClements, & Havstad, 2014). One way of ensuring these conditions is to cultivate and foster a research data infrastructure or cyberinfrastructure (Florio & Sirtori, 2016), which aims to meet the needs of the research community for democratic access to digital resources and collaborative environments around common practices (Atkins, 2003). A cyberinfrastructure includes high performance computing (HPC) and use of large shared data storage, a platform or stack of services that provides methods for leveraging those physical resources, and a community of people and institutes that manage these resources in a sustainable, secure, collaborative, and interoperable way (Goff et al., 2011).

2 | COLOMBIA'S BIODIVERSITY FOSTERING SOCIOECONOMIC GROWTH

Colombia's topography and location near the equator make it a highly biodiverse country (Rangel-Ch, 2015). The country is one of the 17 "megadiverse countries" in the world according to the United Nations Environment Programme (UNEP). Colombia has suffered an expensive internal conflict for five decades, which was only recently alleviated through a peace agreement in late 2016 (Baptiste et al., 2016). Lack of stability and limited opportunities in at least half of the country, particularly remote rural regions, have resulted in evident negative socioeconomic and ecological impacts (Baumann & Kuemmerle, 2016).

The "Colombia BIO" programme lead by the Colombian Research Council (*Colciencias*) is seeking to make sustainable use of this natural capital to drive the growth of the Colombian bioeconomy, social equality, and a long-lasting peace (Sierra et al., 2017). In "Colombia BIO"'s expeditions, large amounts of data about Colombia's ecosystems are being collected, including novel biodiversity in regions that were previously unexplored due to the internal conflict (Gonzalez, Arenas, Tovar, Pulido, & Tenorio, 2017). As 2019, Colombia is one of the 11 country funders of the "Earth Biogenome Project" (EBP; earthbiogenome.org). The EBP "can be viewed as infrastructure for the new biology" that aims to sequence, catalogue, and characterize the genomes of all known eukaryotes to inform ecosystems preservation under the growing impacts from climate change and overexploitation (Lewin et al., 2018). The EBP consortium in Colombia is led by the University of Los Andes and "BRIDGE Colombia" (Prof F. Di Palma, personal communication).

The capacity to share and analyze this information needs to keep pace with the wealth of information gleaned from these new and upcoming explorations (Canhos et al., 2015). To date, the national catalogue of Colombian biodiversity (SiB Colombia) (Abud et al., 2017) includes 7,848 endemic species and around 10% of all known species. Researchers and policymakers need to be provided with comprehensive evidence to inform evidence-supported decisions on biodiversity management and protection.

For that, the Essential biodiversity variables (EBVs) define a minimum common set of data about taxa (distribution, genome, phenome, traits, ecological interactions, etc.) including their environmental and evolutionary context (La Salle, Williams, & Moritz, 2016); although EBVs' practical implementation remains a challenge (Kissling et al., 2018; Pereira et al., 2013).

3 | COLOMBIA'S ASSESSMENT BY "C3BIODIVERSIDAD": INTRODUCING A REFERENCE FRAMEWORK FOR EMERGING COUNTRIES

"C3biodiversidad," the Colombian Cyberinfrastructure Consortium for Biodiversity, aims to develop a research cyberinfrastructure in Colombia, particularly for the analysis of biodiversity data, through the sharing of computing resources already available in the country, promoting new resources under "open access" incentives, and building skilled human capital capable of operating these resources in the long-term. Here (Table S1), we provide a summary of the analysis of Colombia's internal and external strengths and weaknesses (SWOT) for building a local cyberinfrastructure organized into four subjects (following Sections 4 to 7). These conclusions are an output of the "C3biodiversidad workshop" held in Bogota, Colombia in June 2018. The workshop included 36 experts from 16 leading Colombian institutions and a group of international facilitators and experts that represented a fair distribution of interest groups. These discussions also highlighted the need for coordinating with existing centers of excellence in the country, tapping into successful initiatives in the region, and leveraging on existing international open source resources and projects. Building on these conclusions, we have identified key priorities (Table 1) and developed a reference framework (Figure 1) for building cyberinfrastructure. While the conclusions reflect Colombia's environment at the time of the workshop, we believe these can be applied to other middle and upper-middle income countries. The four key priorities are discussed in the following sections.

TABLE 1 Four priorities for building cyberinfrastructure in emerging countries. These priorities should be developed in a coordinated manner by local innovators. For each priority, possible interventions are also included as examples

Priorities	Vision	Mission	Objectives	Open SC. Framework*	Interventions
Improve the provisioning and availability of physical infrastructure	Accelerate data-intensive scientific research.	Organically build a distributed sustainable cooperative computational platform.	• Explore science. • Facilitate skills sharing.	• Open research infrastructures • Enabling e-infrastructures	• Survey the resources needs and availability. • Facilitate physical connectivity among institutions. • Formalize an advisory community about physical computational resources. • Implement a recognition scheme for resources providers in evaluations and panels.
Grow training in scientific data analysis for users and providers	A generation skilled in scientific data analysis.	Promote a coordinated accessible programme of training in scientific data analysis.	• Coordinate advanced tailored training. • Facilitate institutional cooperation.	"Cross-cutting issue"	• Develop a national online network of trainers and communities in data analysis. • Coordinate with global and regional training networks such as GLOBET. • Support hosting of international trainers. • Support staff exchanges from smaller to larger institutions within the country.
Develop and enforce a national policy for research data	Data-supported decision-making.	Develop and enforce a national policy for research data.	• Incentive excellence in research. • Facilitate access to research data	• Open research data • Open licenses and IPR	• Involve stakeholders in new policy design. • Require open access to taxpayer-funded research. • Promote an institution to coordinate repositories and databases within the country. • Reward researchers involved in data partnerships.
Engage diverse stakeholders in research projects and funding planning	A society highly involved and interested in science and technology.	Develop transversal research schemes that reward stakeholder's engagement.	• Support private–public partnerships. • Support third-sector involvement. • Promote multidisciplinarity	• Open research agenda setting • Open funding mechanisms • Open metrics • Citizen science	• Require engagement plan annexed to research projects. • Implement funding calls for public–private projects. • Extend the role of research-support offices. • Catalogue and disseminate networking opportunities.

*Based on the conceptual open-science (OS) framework defined by the Colombian Research Council *Colciencias* (Colciencias, 2018).

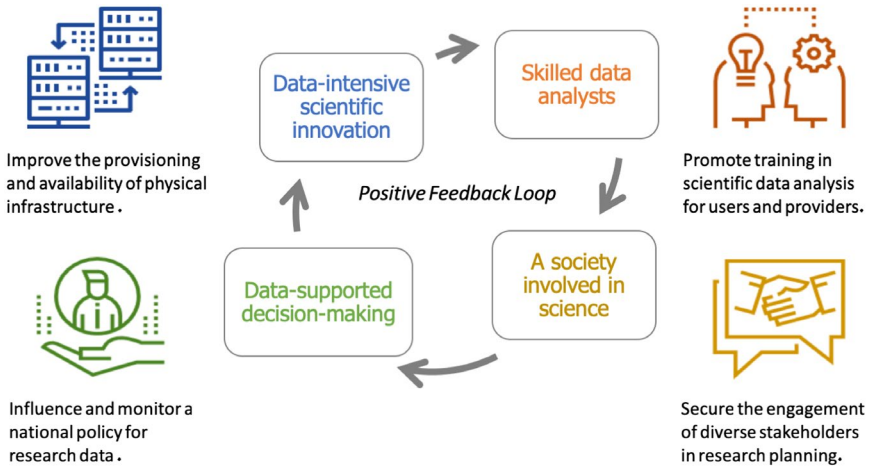

FIGURE 1 A reference framework consisting of four priorities to facilitate the socioeconomic growth in emerging countries through innovation by developing a research cyberinfrastructure

4 | IMPROVING THE PROVISIONING AND AVAILABILITY OF PHYSICAL DATA INFRASTRUCTURE

Biodiversity cyberinfrastructures increase data access and reusability, and also support education and effective public policies. To balance the potential costs in the context of the scientific benefits, the research community often self-organizes to identify the broad-scale questions that require large data-driven analysis that can only be addressed by expensive infrastructure, which is then funded by research councils usually on the condition to be shared as a community resource.

The main challenge Colombia and other middle and upper-middle countries currently face is their limited access to computational capacity and physical connectivity between research institutions. This technological gap is mostly the result of limited funding, the high cost of foundational infrastructure, inconsistent interest from multinational vendors, and short-term strategic planning. As a result, key academic and industrial institutions prioritise limiting uncertainty, unforeseen overheads, and imported commodities. Still, major universities and centers in Colombia and other emerging countries have access to HPC infrastructure (Cazar, 2018). However, these infrastructures are primarily, and usually exclusively, implemented to meet the internal needs of the host institution.

It is a priority to deploy high-performance computational platforms in the institutions of a country as a requirement to accelerate research and skills training. We believe the best option is progressively integrating into increasing orders of complexity existing resources under a fair-sharing policy that prioritises the host institution while promoting sharing new computational and data storage capacities through capital investments and incentives. Infrastructures require substantial financial investments in the hardware itself, physical space, environment control, management, and maintenance. For example, the *CyVerse* cyberinfrastructure is leveraging on the considerable investment from the USA's National Science Foundation (NSF) (Goff et al., 2011).

Distributed infrastructures are composed of multiple independent and distributed resources that act as one, often with resources provided by different institutions (Towns et al., 2014), so the initial costs and complexity are distributed. These are usually rolled out in stages in increasing degree of decentralization (Chaterji et al., 2017).

A federation of heterogeneous computing resources as the kind proposed needs to address two managerial requirements in order to be successful. Firstly, because most institutions want to retain the right to define their own policies on data management and execution priorities, the system must guarantee that users can access each resource at the right level of privilege. As a result, a distributed system typically "authenticates, authorises, and accounts" (AAA system) the user for each individual system in a centralized server. Secondly, when computational resources and data are dispersed in storage locations among participating organizations, end users should be relieved of the complexities associated with negotiating access rights with individual organizations, moving data back and forth, or porting programs to process the data (Langmead & Nellore, 2018). Technical software solutions for example, data management middleware such as the open source *iRODS* software (Rajasekar et al., 2010), workflow software and virtual machines (Boettiger, 2015; Köster & Rahmann, 2012) provide tested options for data federation, data replication, quota management, and access control etc.

A successful precedent of distributed high-performance computational platform is the Iberian-American Network for High-Performance Computing (RICAP, 2017–2020). RICAP's resources are distributed across 11 sites in various Latin-American countries, which are connected through RedClara, the network of Latin America's academic networks (Cazar, 2018). The existence in many emerging countries of state-sponsored high-speed academic-network providers (*Red Nacional Académica de Tecnología Avanzada*, RENATA, in the case of Colombia) is key to facilitate the necessary physical connectivity between institutions. However, our SWOT analysis highlighted that Colombian research institutions actively use the connectivity services from private providers too (Table S1).

5 | IMPLEMENTING A NATIONAL FAIR POLICY FOR RESEARCH DATA MANAGEMENT

The absence of a comprehensive policy that regulates and enforces access to research data restrains research. It is a priority to develop

and implement a national policy for research data that regulates the access, processing and sharing of data in a standardized way. This would facilitate data-supported decision-making, as well as scientific excellence and innovation. In the case of biodiversity, the national implementations on "Access and Benefit Sharing" of genetic resources, designed to give greater control over the natural capital, have also generated regulatory regimes fraught with unintended consequences, this is not exclusive to Colombia (Prathapan et al., 2018; Wight, 2019). In Colombia and other emerging countries, there are well-developed policies that regulate other data types, such as Government (e-gov) and personal data that can serve as examples to develop research data policy (Sanabria, Pliscoff, & Gomes, 2014). We recommend requiring open access to taxpayer-funded research, including both generated data and research publications, as recommended by the Organisation for Economic Co-operation and Development (OECD) (Arzberger et al., 2004). When funding is a limiting factor, policy needs to maximize return on investment in data generation.

Good data management is not a goal in itself, but rather is the key conduit guaranteeing experimental reproducibility (Baker, 2016) and maximizing return on investment in data generation by facilitating its reuse by third parties. Four foundational principles, findability, accessibility, interoperability, and reusability (FAIR) usually guide good data management practices among producers and publishers (Wilkinson et al., 2016). In Colombia, *Colciencias* has recently published its vision to promote an "open science" in the country based on the FAIR principles (Colciencias, 2018). Significant challenges to implementing data management arise from the size and complexities of modern scientific collaboration (Borgman et al., 2012). Still, when psychology researchers were asked to rank barriers to data sharing, technological barriers (such as "My dataset is too big" or "There is no suitable repository to share my data") were at the bottom of the list (Houtkoop et al., 2018). Similar results were obtained in other disciplines (Van den Kaye, Bruce, & Fripp, 2017; Eynden et al., 2016), or in the specific case of Colombian researchers (OCyT, 2017).

Data sharing can be incentivized by normative pressure, for example through a strong centralized information system or due to requirements of funding agencies and journals to release research data at the time of publication or end of funding (Wolkovich, Regetz, & O'connor MI., 2012). In large projects, funding agencies and international directorates will need to work together in joint initiatives to overcome cultural barriers and geopolitical constraints among countries (Vargas et al., 2012). However, regardless of journal or funder requirements, data are routinely shared in some scientific fields as a result of a cultural shift, scholarly altruism, and peer approval (Kim & Stanton, 2012; OCyT, 2017). Also, data sharing can be promoted by recognizing those who analyze it as creative collaborators in need of career paths (Chang, 2015). Highlighting and disseminating specific research communities and projects that follow standards, curation and preservation approaches can serve as showcases (Canhos et al., 2015; Sanabria et al., 2014). For example, SIB Colombia was rewarded as the best "open science initiative" in the country in 2017 by Colciencias. Further interventions in this area include creating the figure of "Data Champions" (volunteers who advise researchers in their institutions on good research data management and promote FAIR guidelines) and promoting a model where institutional repositories would coexist with a centralized national data management repository.

6 | GROWING TRAINING IN SCIENTIFIC DATA ANALYSIS FOR USERS AND PROVIDERS

Skilled labor emigration and limited advanced training opportunities for new recruits are constant risks in middle and upper-middle countries (O'Mahony, Robinson, & Vecchi, 2008). So, it is a priority to design and promote a coordinated programme of training in scientific data analysis tailored to different career levels, as well as providing opportunities for career development, to address "brain drain." The demand for training is high, our analysis of Colombia's situation evidenced that opportunities for coordinated training between strong groups have not been fully explored, and internships and visits between groups are uncommon. Possible interventions include providing technical skills to experts in data analysis, coordinating the training offered in the country, engaging with the global training communities and funding visits from international trainers and staff exchanges. The amount of data generated by high-throughput experimental technologies has increased the demand for scientists involved in research to acquire a minimum set of capabilities in bioinformatics to effectively communicate with bioinformaticians (Tan, Lim, Khan, & Ranganathan, 2009; Welch et al., 2014). The Global Organisation for Bioinformatics Learning, Education and Training (GOBLET) surveys provide "perspectives on the current status of training gaps" and evidence that "the need for bioinformatics training is both real and urgent, and requires worldwide solutions" (Attwood et al., 2015).

Running effective courses and workshops means having tailored teaching materials and instructors trained in how to teach students who may come from different backgrounds and have different goals. Not surprisingly, the completion rate for self-paced Massive Open Online Courses (MOOCs) is less than 10% (Jordan, 2014). However, trainers are available in Colombia and equivalent countries. For example, there is an academic network in Colombia focused on bioinformatics, as well as a biannual national bioinformatics conference, which is often organized in collaboration with other scientific societies. Another key strength is the availability of graduate system administrators and developers; formal training is available through at least four M.Sc. programmes in bioinformatics, data science, or computational biology, as well as several in computational sciences. On the one hand, Train-the-Trainers (TTT) workshops, where future instructors are equipped with practical skills to effectively teach, are a cost-effective way to prepare instructors (Pfund et al., 2015; Via et al., 2017). On the other hand, the "keep training local but act to deliver and develop training materials globally" motto highlights how a community might break down the effort of producing training

materials in a modular way (Williams & Teal, 2017). This decentralized approach allows training to become more accessible to more people while "responding at scale to rapidly evolving science" (Teal et al., 2015). For example, software Carpentry and Data Carpentry lessons are developed collaboratively on Github by volunteers.

7 | SECURING THE ENGAGEMENT OF DIVERSE STAKEHOLDERS IN PLANNING

We believe it is a priority securing the engagement of a diverse range of stakeholders in research planning, and particularly in cyberinfrastructure planning and execution. Researchers are the driving force in the innovation process, and they will only engage in the cyberinfrastructure if they perceive the cyberinfrastructure as a way to ease data management and analysis. There is consequently a need to survey a priori the needs of the community (Cutcher-Gershenfeld et al., 2016; Nativi, Craglia, & Pearlman, 2013). For example, the DataONE cyberinfrastructure (https://www.dataone.org/) (Michener et al., 2012) used four "participatory user-centred" workshops during its inception. The responses from the survey on "open science and policy" to 564 Colombian researchers (OCyT, 2017) help frame and assert the need and receptivity to a cyberinfrastructure as proposed in Colombia. As summarized in Table S2, Colombian researchers' priorities were to "develop strategies and tools" (91%), "promote skills exchange" (83%), "design incentives" (80%), and "support best practices" (78%); and also found "data availability" (72%), "digital technology and capacities" (62%), and "new ways for dissemination (59%) and collaboration (55%)" as courses of action (Table S2).

The workshop results also proposed promoting private–public partnerships and extending the involvement of the third sector (non-profit associations, charities, cooperatives, etc.) in research. While researchers are the driving force in the innovation process, the environment where each researcher works (industry, academia, nonprofit, general public, or government) frames how researchers can conduct that research. Our analysis in Colombia highlighted that there is a limited number of initiatives to engage stakeholders in research and a variable interest in research from different sectors. Partnerships between industry, third sector, government and academia appear to be more established in the agricultural and environmental sectors, for example. We identified the following three positive recent initiatives in Colombia: 1. Specific research public funding opportunities involving industry; 2. a new research funding system from the regions to promote regional redistribution; and 3. increasing international investment after Colombia's access to the OECD and the peace agreement process.

Finally, secondary stakeholders (citizens, educators, librarians, policymakers, funding officers, editors, professional societies, etc.) have their particular interests and priorities, and consequently a say in planning. When asked about the impact of open science on society, researchers in Colombia highlighted the mutual benefits of improving the social awareness, reproducibility and general efficiency of science (OCyT, 2017).

ACKNOWLEDGMENTS

The authors would like to acknowledge the support from the UK Research and Innovation (UKRI) Global Challenges Research Fund (GCRF) GROW Colombia grant via the UK's Biotechnology and Biological Sciences Research Council (BB/P028098/1), as well as from Colciencias Colombia BIO project and the British Council in Colombia. This publication builds on the analysis from a panel of experts at the Colombian Science Council (Colciencias) in Bogota, Colombia on 16-18 June 2018. As a result, we would like to acknowledge the contributions of Alejandro Caro, AGROSAVIA; Andrés Pinzón Velasco, National University of Colombia; Camilo Corchuelo Rodríguez, Santo Tomás University; Cesar Orlando Díaz, Jorge Tadeo Lozano University; Daniel Fernando López, Humboldt Institute; Dany Molina, Colombia's Center for Bioinformatics and Computational Biology (BIOS); Diego Rincón, Catholic University of Colombia; Gastón Lyons, University of Los Andes; Jaime Erazo, Earlham Institute; John Jaime Riascos, CENICAÑA; Jorge William, Colombia's Center for Bioinformatics and Computational Biology (BIOS); Juan Manuel Anzola, Corpogen; Laura Natalia González García, University of Los Andes; Leroy Mwanzia, International Center for Tropical Agriculture (CIAT); Luz Miriam Díaz, National Academic Network of Advanced Technology of Colombia (RENATA); María Camila Martínez, CENICAÑA; Patricia Jaramillo, National Academic Network of Advanced Technology of Colombia (RENATA); Paula Reyes, AGROSAVIA; Raúl Ramos Pollán, University of Antioquia; Romain Guyot, Autonomous University of Manizales; Tomás Viloria Lagares, University of Los Llanos; and Yesid Cuesta Astroz, University of Antioquia.

AUTHOR CONTRIBUTIONS

JdV, RPD, WH, and FdP conceived and financed this work. EB-H, JD, and JPM-R organized the stakeholders' workshop where the data were collected. JdV, RPD, GJH, AM, MM-T, and NF-F coached the workshop and facilitated the data collection. All authors participated and contributed to the "C3biodiversidad" workshop in Bogota upon the conclusions of which this manuscript builds on, particularly DE, MAC-A, NEA-S, JDP-D, JC-A, and AVCR analyzed and interpreted data for the SWOT analysis. All authors contributed to the discussion of the structure of the manuscript. JdV, RPD, and NF-F drafted the article with contributions by all authors at different stages. All the authors revised and approved the final version.

ORCID

Jose J. De Vega https://orcid.org/0000-0003-2847-5158
Robert P. Davey https://orcid.org/0000-0002-5589-7754
Jorge Duitama https://orcid.org/0000-0002-9105-6266
Dairo Escobar https://orcid.org/0000-0001-8327-8670
Marco A. Cristancho-Ardila https://orcid.org/0000-0001-9115-4117
Graham J. Etherington https://orcid.org/0000-0002-5003-1425
Alice Minotto https://orcid.org/0000-0002-1670-1675
Nelson E. Arenas-Suarez https://orcid.org/0000-0002-7665-8955

Javier Correa-Alvarez https://orcid.org/0000-0001-9009-823X
Anyela V. Camargo Rodriguez https://orcid.org/0000-0003-4277-0356
Wilfried Haerty https://orcid.org/0000-0003-0111-191X
Juan P. Mallarino-Robayo https://orcid.org/0000-0002-7424-3000
Emiliano Barreto-Hernandez https://orcid.org/0000-0002-5968-6794
Monica Muñoz-Torres https://orcid.org/0000-0001-8430-6039
Narcis Fernandez-Fuentes https://orcid.org/0000-0002-6421-1080
Federica Di Palma https://orcid.org/0000-0002-4394-0102

REFERENCES

Abud, M., Agudelo, E., Aguilar-Cano, J. R., Alvarez Aristizábal, A., Andrade, Á., Andrade-Pérez, G. I., … Dueñas, J. A. (2017). *Biodiversidad 2017: Estado y tendencias de la biodiversidad continental de Colombia*. Bogotá, Colombia: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt.

Annan, K. (2003). A challenge to the world's scientists. *Science*, *299*, 1489. https://doi.org/10.1126/science.299.5612.1485

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., … Wouters, P. (2004). An international framework to promote access to data. *Science*, *303*, 1777–1778. https://doi.org/10.1126/science.1095958

Atkins, D. (2003). Revolutionizing science and engineering through cyberinfrastructure. Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure. NSF, WA, USA. https://doi.org/10150/106224

Attwood, T. K., Bongcam-Rudloff, E., Brazas, M. E., Corpas, M., Gaudet, P., Lewitter, F., … Van, C. G. (2015). GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training. *PLoS Computational Biology*, *11*, e1004143. https://doi.org/10.1371/journal.pcbi.1004143

Baker, M. (2016). Reproducibility: Seek out stronger science. *Nature*, *537*, 703–704. https://doi.org/10.1038/nj7622-703a

Baptiste, B., Pinedo-Vasquez, M., Gutierrez-Velez, V. H., Andrade, G. I., Vieira, P., Estupiñán-Suárez, L. M., … Lee, T. M. (2016). Greening peace in Colombia. *Nature Ecology & Evolution*, *1*, 102. https://doi.org/10.1038/s41559-017-0102

Baumann, M., & Kuemmerle, T. (2016). The impacts of warfare and armed conflict on land systems. *Journal of Land Use Science*, *11*, 672–688. https://doi.org/10.1080/1747423X.2016.1241317

Boettiger, C. (2015). An introduction to Docker for reproducible research, with examples from the R environment. *ACM SIGOPS Operating Systems Review*, *49*, 71–79. https://doi.org/10.1145/2723872.2723882

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work (CSCW)*, *21*, 485–523. https://doi.org/10.1007/s10606-012-9169-z

Canhos, D. A., Sousa-Baena, M. S., de Souza, S., Maia, L. C., Stehmann, J. R., Canhos, V. P., … Peterson, A. T. (2015). The importance of biodiversity e-infrastructures for megadiverse countries. *PLoS Biology*, *13*, e1002204. https://doi.org/10.1371/journal.pbio.1002204

Chang, J. (2015). Core services: Reward bioinformaticians. *Nature*, *520*, 151–153. https://doi.org/10.1038/520151a

Chaterji, S., Koo, J., Li, N., Meyer, F., Grama, A., & Bagchi, S. (2017). Federation in genomics pipelines: Techniques and challenges. *Briefings in Bioinformatics*, *20* (1), 235–244. https://doi.org/10.1093/bib/bbx102

Colciencias (2018). Lineamientos Para Una Política De Ciencia Abierta En Colombia, Documento de Política Nacional de Ciencia, Tecnología e Innovación número 1801. Colciencias, Bogotá, Colombia.

Cutcher-Gershenfeld, J., Baker, K. S., Berente, N., Carter, D. R., DeChurch, L. A., Flint, C. G., … Kirkpatrick, C. (2016). Build it, but will they come? A geoscience cyberinfrastructure baseline analysis. *Data Science Journal*, *15*(8), 1–14. https://doi.org/10.5334/dsj-2016-008

Florio, M., & Sirtori, E. (2016). Social benefits and costs of large scale research infrastructures. *Technological Forecasting and Social Change*, *112*, 65–78. https://doi.org/10.1016/j.techfore.2015.11.024

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., … Stanzione, D. (2011). The iPlant collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science*, *2*, 34. https://doi.org/10.3389/fpls.2011.00034

Gonzalez, M. A., Arenas, H., Tovar, E., Pulido, P., & Tenorio, E. (2017). Colombia BIO: Discovering biodiversity in post-conflict territories in Colombia. *Genome*, *60*, 938–939.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, *34*, 135–174. https://doi.org/10.1111/dpr.12142

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, *1*, 70–85. https://doi.org/10.1177/2515245917751886

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, *15*(1), https://doi.org/10.19173/irrodl.v15i1.1651

Kaye, J., Bruce, R., & Fripp, D. (2017). Establishing a shared research data service for UK universities. *Insights the UKSG Journal*, *30*(1), 59–70. https://doi.org/10.1629/uksg.346

Kim, Y., & Stanton, J. M. (2012). Institutional and individual influences on scientists' data sharing practices. *Journal of Computational Science Education*, *3*, 47–56. https://doi.org/10.22369/issn.2153-4136/3/1/6

Kissling, W. D., Walls, R., Bowser, A., Jones, M. O., Kattge, J., Agosti, D., … Guralnick, R. P. (2018). Towards global data products of Essential Biodiversity Variables on species traits. *Nature Ecology & Evolution*, *2*, 1531. https://doi.org/10.1038/s41559-018-0667-3

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*, 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

La Salle, J., Williams, K. J., & Moritz, C. (2016). Biodiversity analysis in the digital era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150337. https://doi.org/10.1098/rstb.2015.0337

Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, *19*, 208. https://doi.org/10.1038/nrg.2017.113

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., … Zhang, G. (2018). Earth BioGenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*, 4325–4333. https://doi.org/10.1073/pnas.1720115115

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., … Vieglais, D. A. (2012). Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, *11*, 5–15. https://doi.org/10.1016/j.ecoinf.2011.08.007

Mocskos, E., Barrios, H., Carlos, J., Castro, H., Ramirez, D. C., Nesmachnow, S., & Mayo-Garcia, R. (2018). Boosting advanced computational applications and resources in Latin America through

collaboration and sharing. *Computing in Science & Engineering*, *20*, 39. https://doi.org/10.1109/MCSE.2018.03202633

Nativi, S., Craglia, M., & Pearlman, J. (2013). Earth science infrastructures interoperability: The brokering approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(3), 1118–1129. https://doi.org/10.1109/JSTARS.2013.2243113

O'Mahony, M., Robinson, C., & Vecchi, M. (2008). The impact of ICT on the demand for skilled labour: A cross-country comparison. *Labour Economics*, *15*, 1435–1450. https://doi.org/10.1016/j.labeco.2008.02.001

OCyT (2017). Estudio para identificar conocimientos, capacidades, percepciones y experiencias de los investigadores del país frente a la ciencia abierta. Observatorio Colombiano de Ciencia y Tecnología (OCyT), Bogotá, Colombia. http://repositorio.colciencias.gov.co/handle/11146/21720

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., … Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*, 277–278. https://doi.org/10.1126/science.1229931

Peters, D. P., Loescher, H. W., SanClements, M. D., & Havstad, K. M. (2014). Taking the pulse of a continent: Expanding site-based research infrastructure for regional-to continental-scale ecology. *Ecosphere*, *5*, 1–23. https://doi.org/10.1890/ES13-00295.1

Pfund, C., Spencer, K. C., Asquith, P., House, S. C., Miller, S., & Sorkness, C. A. (2015). Building national capacity for research mentor training: An evidence-based approach to training the trainers. *CBE—Life Sciences Education*, *14*, ar24. https://doi.org/10.1187/cbe.14-10-0184

Prathapan, K. D., Pethiyagoda, R., Bawa, K. S., Raven, P. H., Rajan, P. D., & Countries 172 Co-Signatories from 35. (2018). When the cure kills—CBD limits biodiversity research. *Science*, *360*, 1405–1406. https://doi.org/10.1126/science.aat9844

Rajasekar, A., Moore, R., Hou, C., Lee, C. A., Marciano, R., de Torcy, A., … Gilbert, L. (2010). iRODS primer: Integrated rule-oriented data system. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *2*, 1–143.

Rangel-Ch, J. O. (2015). La biodiversidad de Colombia: Significado y distribución regional. *Revista De La Academia Colombiana De Ciencias Exactas, Físicas Y Naturales*, *39*, 176–200.

Sanabria, P., Pliscoff, C., & Gomes, R. (2014). E-government practices in South American countries: Echoing a global trend or really improving governance? The experiences of Colombia, Chile, and Brazil. In M. Gascó-Hernández (Ed.), *Open Government* (pp. 17–36). New York: Springer. https://doi.org/10.1007/978-1-4614-9563-5_2

Sierra, C. A., Mahecha, M., Poveda, G., Álvarez-Dávila, E., Gutierrez-Velez, V. H., Reu, B., … Skowronek, S. (2017). Monitoring ecological change during rapid socio-economic and political transitions: Colombian ecosystems in the post-conflict era. *Environmental Science & Policy*, *76*, 40–49. https://doi.org/10.1016/j.envsci.2017.06.011

Tan, T. W., Lim, S. J., Khan, A. M., & Ranganathan, S. (2009). A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the "-omics" era. *BMC Genomics*, *10*, S36. https://doi.org/10.1186/1471-2164-10-S3-S36

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., & Pawlik, A. (2015). Data carpentry: Workshops to increase data literacy for researchers. *International Journal of Digital Curation*, *10*, 135–143. https://doi.org/10.2218/ijdc.v10i1.351

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., … Wilkins-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering*, *16*, 62–74. https://doi.org/10.1109/MCSE.2014.80

Van den Eynden, V., Knight, G., Vlad, A., Radler, B., Tenopir, C., Leon, D., … Corti, L. (2016). Survey of Wellcome researchers and their attitudes to open research. Technical Report. Wellcome Trust, London, UK. https://doi.org/10.6084/m9.figshare.4055448.v1

Vargas, R., Loescher, H. W., Arredondo, T., Huber-Sannwald, E., Lara-Lara, R., & Yépez, E. A. (2012). Opportunities for advancing carbon cycle science in Mexico: Toward a continental scale understanding. *Environmental Science & Policy*, *21*, 84–93. https://doi.org/10.1016/j.envsci.2012.04.003

Via, A., Attwood, T. K., Fernandes, P. L., Morgan, S. L., Schneider, M. V., Palagi, P. M., … Tractenberg, R. E. (2017). A new pan-European Train-the-Trainer Programme for bioinformatics: Pilot results on feasibility, utility and sustainability of learning. *Briefings in Bioinformatics*, *20*(2), 405–415. https://doi.org/10.1093/bib/bbx112

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., & Schneider, M. V. (2014). Bioinformatics curriculum guidelines: Toward a definition of core competencies. *PLOS Computational Biology*, *10*, e1003496. https://doi.org/10.1371/journal.pcbi.1003496

Wight, A. J. (2019). In Colombia, biodiversity researchers seek relief from regulatory red tape. *Science News*. https://doi.org/10.1126/science.aax1404

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Williams, J. J., & Teal, T. K. (2017). A vision for collaborative training infrastructure for bioinformatics. *Annals of the New York Academy of Sciences*, *1387*, 54–60. https://doi.org/10.1111/nyas.13207

Wolkovich, E. M., Regetz, J., & O'Connor, M. I. (2012). Advances in global change research require open science by individual researchers. *Global Change Biology*, *18*, 2102–2110. https://doi.org/10.1111/j.1365-2486.2012.02693.x

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth*, *10*, 13–53. https://doi.org/10.1080/17538947.2016.1239771

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: De Vega JJ, Davey RP, Duitama J, et al; the Colombian Cyberinfrastructure Consortium for Biodiversity. Colombia's cyberinfrastructure for biodiversity: Building data infrastructure in emerging countries to foster socioeconomic growth. *Plants*, *People*, *Planet*. 2019;00:1–8. https://doi.org/10.1002/ppp3.10086